

A fast iteration method for solving elliptic problems with quasiperiodic coefficients

Boris N. Khoromskij*

Sergey I. Repin**

Dedicated to 70th jubilee of Prof. Yu. A. Kuznetsov

Abstract

The paper suggests a preconditioning type method for fast solving of elliptic equations with oscillating quasiperiodic coefficients A_ϵ specified by the small parameter $\epsilon > 0$. We use an iteration method generated by an elliptic operator, associated with a certain simplified (e.g., homogenized) problem. On each step of this procedure it is required to solve an auxiliary elliptic boundary value problem with non-oscillating coefficients A_0 . All the information related to complicated coefficients of the original differential problem is encompasses in the linear functional, which forms the right hand side of the auxiliary problem. Therefore, explicit inversion of the original operator associated with oscillating coefficients is avoided. The only operation used instead is multiplication of the operator by a vector (vector function), which can be efficiently performed due to the low-rank QTT tensor operations with the rank parameter controlled by the given precision $\delta > 0$ independent on the parameter ϵ . In the first part of the paper, we establish sufficient conditions that guarantee convergence of the iteration method and deduce explicit estimates of the contraction factor, which are expressed in terms of A_ϵ and A_0 . Moreover, we deduce two-sided a posteriori error estimates that do not use A_ϵ^{-1} and provide guaranteed two sided bounds of the distance to the exact solution of the original problem for any step of the iteration process. The second part is concerned with realisations of the iteration method. For a wide class of oscillating coefficients, we obtain sharp QTT rank estimates for the stiffness matrix in tensor representation. In practice, this leads to the logarithmic complexity scaling of the approximation and solution process in both the FEM grid-size, and $O(|\log \epsilon|)$ cost in terms of ϵ . Numerical tests in 1D confirm the logarithmic complexity scaling of our method applied to a class of complicated quasiperiodic coefficients.

AMS Subject Classification: 65F30, 65F50, 65N35, 65F10

Key words: lattice-structured and quasi-periodic systems, a posteriori estimates, tensor numerical methods, quantized tensor approximation, block-structured matrices, preconditioning.

*Max Planck Institute for Mathematics in the Sciences, Inselstr. 22-26, D-04103 Leipzig, Germany (bokh@mis.mpg.de)

**V.A. Steklov Institute of Mathematics, 191011, Fontanka 27, St.Petersburg, Russia (repin@pdmi.ras.ru)

1 Introduction

Partial differential equations with oscillating coefficients often arise in various models in natural sciences, including quantum chemistry and material sciences, as well as in engineering applications. Numerical analysis of problems with periodical coefficients is often performed by geometric homogenization methods, which provide efficient approximations of structures with very large amount of cells of periodicity (see, e.g., [1, 2, 3, 5, 4]). We consider a wider class of problems where either the amount of cells is significant but not large enough to ignore modeling errors generated by homogenized models or periodicity has a more complicated form. Numerical analysis of such problems is faced with several challenging problems. The main three of them are as follows: (a) creation of a robust numerical method able to construct a sequence converging to the exact solution by means of using finite element approximations on regular (quasiregular) meshes; (b) guaranteed a posteriori estimates of the distance between the exact solution of a boundary value problem with highly oscillating coefficients and an approximation; (c) construction an efficient solver based on suitable preconditioning of the respective discrete system.

In this paper, we suggest an approach that solves (a)–(c) for a class of elliptic problems with quasi-periodic coefficients. We discuss the basic ideas with the paradigm of the model second-order elliptic problem, but it is clear that they can be extended to many other elliptic and parabolic type equations with quai-periodic coefficients. Consider the problem

$$-\operatorname{div}(A_\epsilon(x)\nabla u_\epsilon) = f \quad \text{in } \Omega, \quad u_\epsilon = 0 \quad \text{on } \partial\Omega, \quad (1.1)$$

where $f \in L^2(\Omega)$, $\Omega = (0, 1)^d$ ($d = 1, 2, 3$), with homogeneous Dirichlet boundary conditions, where a small parameter $\epsilon > 0$ is a small parameter characterizing oscillations, and $A_\epsilon(x)$ is a matrix with quasiperiodic coefficients. We assume that $A_\epsilon \in L^\infty(\Omega, \mathbb{M}_{\text{sym}}^{d \times d})$ (here and later on $\mathbb{M}_{\text{sym}}^{d \times d}$ denotes the set of symmetric $d \times d$ - matrices) and

$$\lambda_\ominus^\epsilon |\zeta|^2 \leq A_\epsilon(x) \zeta \cdot \zeta \leq \lambda_\oplus^\epsilon |\zeta|^2, \quad \forall \zeta \in \mathbb{R}^d, \quad x \in \Omega, \quad (1.2)$$

where λ_\ominus^ϵ is a positive constant, so that the problem is well posed and the corresponding generalized solution $u_\epsilon \in H_0^1(\Omega)$ is defined by the relation

$$a_\epsilon(u_\epsilon, w) = (f, w)_\Omega \in H_0^1(\Omega), \quad (1.3)$$

where

$$a_\epsilon(u, w) := \int_\Omega A_\epsilon \nabla u_\epsilon \cdot \nabla w dx \quad \text{and} \quad (f, w)_\Omega := \int_\Omega f w dx.$$

Entries of A_ϵ may depend on x in a very complicated way, see some examples depicted in Fig. 1.1. Therefore, the problem (1.1) may be very difficult from the viewpoint of quantitative analysis. The level of complexity can be roughly estimated by the parameters $\kappa := \frac{\lambda_\ominus^\epsilon}{\lambda_\oplus^\epsilon}$ and ϵ . If both of them are very small, then serious difficulties will arise in approximation methods and in numerical solution of the corresponding linear systems (which may have very large dimensions and huge condition numbers). For such type problems, getting guaranteed and efficient a posteriori error estimates may be a highly difficult problem as well.

Our goal is to justify a numerical method for computing successful approximations of u_ϵ which is based on solving a simpler problem associated with the bilinear form $a_0(u, w) =$

$\int_{\Omega} A_0 \nabla u \cdot \nabla w \, dx$. Coefficients of the matrices A_0 are much more regular than coefficients of A_ϵ and do not have rapid oscillations. It is assumed that A_0 satisfies the condition

$$\lambda_{\ominus}^0 |\zeta|^2 \leq A_0 \zeta \cdot \zeta \leq \lambda_{\oplus}^0 |\zeta|^2 \quad \forall \zeta \in \mathbb{R}^d, \, x \in \Omega \quad (1.4)$$

with positive constants λ_{\ominus}^0 and λ_{\oplus}^0 . Then, there exist positive constants λ_1 and λ_2 such that

$$\lambda_1 A_0 \zeta \cdot \zeta \leq A_\epsilon \zeta \cdot \zeta \leq \lambda_2 A_0 \zeta \cdot \zeta \quad \forall \zeta \in \mathbb{R}^d, \, x \in \Omega. \quad (1.5)$$

The homogenization theory suggests a suitable form of $a_0(u, w)$ for perfectly periodical structures, where Ω is a collection of self-similar cells Π_i^ϵ , $i = 1, 2, \dots, L$ and the cell size ϵ is very small (in comparison with the $\text{diam}(\Omega)$). In this case, for any $x \in \Pi_i^\epsilon$ the matrix is defined by the relation $A_\epsilon(x) := \widehat{A}(y) \in L^\infty(\widehat{\Pi}, \mathbb{M}_{\text{sym}}^{d \times d})$, where $y = \frac{x - \zeta_i}{\epsilon}$, ζ_i is the "cell centre", and y is the cartesian coordinate system associated with the "reference" cell $\widehat{\Pi}$. An approximation of u_ϵ is constructed by a special procedure. First, for $k = 1, 2, \dots, d$ we find the solutions N_k of "cell problems"

$$\text{div}(\widehat{A} \nabla N_k) = (\text{div} \widehat{A})_k \quad \text{in} \quad \widehat{\Pi}, \quad (1.6)$$

which satisfy the the periodic boundary conditions and the mean value condition

$$\{N_k\}_{\widehat{\Pi}} := \frac{1}{|\widehat{\Pi}|} \int_{\widehat{\Pi}} N_k = 0.$$

Then, we define the matrix $A_0 = \left\{ \widehat{A}(I - \nabla \mathbf{N}) \right\}_{\widehat{\Pi}}$, where $\mathbf{N} = \{N_1, N_2, \dots, N_d\}$. The homogenized problem is to find $u_0 \in H_0^1(\Omega)$ such that

$$a_0(u_0, w) = (f, w)_\Omega \quad \forall w \in H_0^1(\Omega), \quad (1.7)$$

where $a_0(u_0, w) := \int_{\Omega} A_0 \nabla u_0 \cdot \nabla w \, dx$. This problem is much simpler than the original one. The function u_0 approximates u_ϵ in a weak sense (see, e.g., [2]),

$$u_\epsilon \rightarrow u_0 \quad \text{in } L^2(\Omega) \quad \text{and} \quad u_\epsilon \rightharpoonup u_0 \quad \text{in } H_0^1(\Omega) \quad \text{for } \epsilon \rightarrow 0.$$

In order to obtain a strongly convergent sequence, the homogenization theory suggests to use approximations with a correction, namely,

$$w_\epsilon^1(x) := u_0(x) - \epsilon \psi^\epsilon(x) N_k \left(\frac{x - x_i}{\epsilon} \right) \frac{\partial u_0(x)}{\partial x_k} \quad \forall x \in \Pi_i^\epsilon, \quad \forall i,$$

where $\psi^\epsilon := \min\{1, \frac{1}{\epsilon} \text{dist}(x, \partial\Omega)\}$ is a cutoff function. Then, optimal a priori convergence rates for the error $u_\epsilon - w_\epsilon^1$ can be proved (e.g., see [2] Rem. 5.13, [3], [5] p.28) if $u_0 \in W^{2,\infty}(\overline{\Omega})$ and $\frac{\partial N_k}{\partial y_j} \in L^\infty(\widehat{\Pi})$. The resultant error estimate reads

$$\|u_\epsilon - w_\epsilon^1\|_{H^1(\Omega)} \leq \widetilde{c} \sqrt{\epsilon}. \quad (1.8)$$

Reconstruction of the flux $A_\epsilon \nabla u_\epsilon$ with the same convergence rate $\sqrt{\epsilon}$, requires solving another periodic problem for the operator $\text{curl } A_0^{-1} \text{curl}$.

In general, above discussed correction techniques may be rather costly and the respective convergence estimates usually require additional assumptions concerning regularity of homogenized solutions. It uses solutions of boundary value problems on the cell $\widehat{\Pi}$ (e.g., (1.6)), which often can be found only approximately and require analysis of effects generated by approximation errors and their influence on the accuracy of a_0 , u_0 , w_ϵ^1 , etc. It should be also noted that the homogenization method provides accurate approximations only for sufficiently small ϵ (this fact follows from the a priori estimate (1.8)). The question on how to efficiently compute accurate approximations if ϵ is small but not "very small" remains open. One possible answer is suggested below.

The approach considered in the paper is applicable to a much wider class of problems than problems with periodically oscillating coefficients. It is valid for quasi-periodic structures (e.g. of the type presented in Fig. 1.1, or periodic systems with defects, see [13]), where homogenization theory cannot be used (see examples in §5). Structures of this type arise in various models in natural sciences and engineering applications (see e.g. [11, 12] for applications in electronic structure calculations), so that getting efficient approximations with guaranteed error bounds is an important problem.

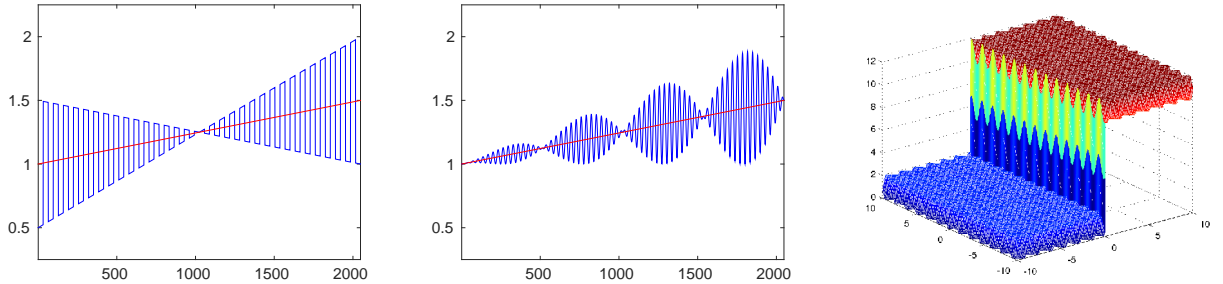


Figure 1.1: Examples of modulated periodic and piecewise periodic coefficients in 1D.

We present a new computational strategy, which is intended to compute efficiently approximate solutions of boundary value problems with periodic and quasi-periodic structures in domains composed of few tensor-product subdomains. This approach is based on the representation (approximation) of all vectors and matrices involved in the computational scheme in the so-called quantized TT (QTT) format [14] such that all matrix-vector operations are implemented approximately via adaptive control the QTT rank parameters. This method allows us to achieve the desired tolerance level regardless of the cell size ϵ , i.e., it does not have limitations of the type (1.8). Under certain assumptions on the tensor structure in coefficients and right-hand side the numerical cost can be estimated by $O(|\log^q \epsilon|)$, where the constant $q > 0$ does not depend on ϵ .

In short, the main ideas behind this approach are as follows. We use a simplified model with much simpler matrix A_0 as the basis of the iteration algorithm (2.6). If A_ϵ is defined by a perfectly regular and highly oscillating structure, then setting A_0 by (1.7) is one possible option. However, there are other options and the choice of A_0 is restricted only by the convergence conditions stated in Theorem 2.1. In other words, we can use any matrix A_0 with simplified or averaged coefficients that coarsely approximate the coefficients of A_ϵ if it provides contraction of the operator T defined by (2.1). In more complicated cases, we can combine averaging, smoothing, and homogenization in different parts of the domain

in dependence of the structure and frequency of oscillations. The possible choice of A_0 is depicted in Fig. 1.1 by red lines (see also examples in Fig. 5.1 and the discussion in Section 4.4).

The structure of A_0 defines the value contraction factor q , which is explicitly estimated a priori. Setting a collection of simplified problems, one can a priori find the problem with minimal q (this amounts solving a simple optimization problem). Next, on each step of the iteration algorithm we have guaranteed two-sided a posteriori error estimates that control the distance to the exact solution (see Section 3).

It is important to outline that all steps of the iteration procedure and error estimates do not require inversion of the matrix generated by the original (complicated) differential operator. This matrix is involved only in multiplication procedures, which can be performed very fast and under in the small storage costs due to QTT tensor operations. This technique is introduced in Section 4, which considers the low-rank quantized tensor approximations [14] arising in the framework of our tensor-based computational scheme. We describe the QTT based preconditioned iteration and present the explicit QTT rank estimates for some particular classes of equation coefficients in 1D. We end up with numerical illustrations demonstrating the fast convergence of preconditioned iteration (geometric convergence rate), as well as the logarithmic scaling of CPU computational time in the grid-size $N = 2^L$. Finally we note that solving differential equations with complicated and rapidly changing coefficients often require special methods and approximations on distorted meshes (see, e.g., [19, 20, 21] and references therein). We believe that modifications of our approach can be also helpful for these cases.

2 The iteration method

2.1 Convergence

We apply the general iteration method (see, e.g., [7]) in order to solve (1.1) with the help of a simpler problem generated by A_0 . Let $v \in V_0$. The functional $\ell_v : V_0 \rightarrow \mathbb{R}$ defined by the relation $\ell_v(w) = a_\epsilon(v, w) - (f, w)_\Omega$ is a linear continuous functional. For any $\rho > 0$ the problem: find $u \in V_0$ such that

$$a_0(u, w) = a_0(v, w) - \rho \ell_v(w) \quad \forall w \in V_0 \quad (2.1)$$

is well posed. Evidently, (2.1) defines a linear bounded operator $T : V_0 \rightarrow V_0$ ($Tv = u$), which is contractive provided that ρ is properly selected. Indeed, select two arbitrary functions $v_1, v_2 \in V_0$ and set $u_1 := T(v_1)$, $u_2 := T(v_2)$. Let $e := v_1 - v_2$, and $\eta := u_1 - u_2$. Then

$$a_0(\eta, w) = a_0(e, w) - \rho \ell_e(w) = a_0(e, w) - \rho a_\epsilon(e, w). \quad (2.2)$$

In view of (2.2), the difference of images is subject to the relation

$$\begin{aligned} \|\eta\|_0^2 &:= a_0(\eta, \eta) = a_0(e, \eta) - \rho a_\epsilon(e, \eta) = \int_{\Omega} (A_0 \nabla e - \rho A_\epsilon \nabla e) \cdot \nabla \eta \, dx \\ &\leq \|\eta\|_0 \left(\int_{\Omega} (A_0 \nabla e - \rho A_\epsilon \nabla e) \cdot (\nabla e - \rho A_0^{-1} A_\epsilon \nabla e) \, dx \right)^{1/2}. \end{aligned} \quad (2.3)$$

Hence,

$$\|\eta\|_0^2 \leq \int_{\Omega} (A_0 - 2\rho A_{\epsilon} + \rho^2 A_{\epsilon} A_0^{-1} A_{\epsilon}) \nabla e \cdot \nabla e \, dx \leq q^2(\rho) \|e\|_0^2, \quad (2.4)$$

where $q^2(\rho) := 1 - 2\rho\lambda_1 + \rho^2 c_{\oplus}$ and c_{\oplus} is the constant in the estimate

$$A_{\epsilon} A_0^{-1} A_{\epsilon} \zeta \cdot \zeta \leq c_{\oplus} A_0 \zeta \cdot \zeta \quad \forall \zeta \in \mathbb{R}^d, \quad x \in \Omega. \quad (2.5)$$

If $0 < \rho < 2\rho_*$, $\rho_* = \frac{\lambda_1}{c_{\oplus}}$ then the quantity $q(\rho)$ is smaller than 1 and (2.4) yields the contraction estimate $\|\eta_0\|_0 \leq q(\rho) \|e\|_0$. Well known results in the theory of fixed points (e.g., see [32]) yield the following result.

Theorem 2.1 *For any $u_0 \in V_0$ and $\rho \in (0, 2\rho_*)$ the sequence $\{u_k\}$ of functions satisfying the relation*

$$a_0(u_{k+1}, w) = a_0(u_k, w) - \rho \ell_{u_k}(w) \quad \forall w \in V_0, \quad (2.6)$$

converges to u_{ϵ} in V and $\|u_k - u_{\epsilon}\|_0 \leq q^k(\rho) \|u_0 - u_{\epsilon}\|_0$ as $k \rightarrow +\infty$.

2.2 Estimates of the contraction parameter

For $\rho = \rho_*$ we find the contraction factor

$$q(\rho_*) = \sqrt{1 - \frac{\lambda_1^2}{c_{\oplus}}}. \quad (2.7)$$

Since $A_0^{-1} A_{\epsilon} \zeta \cdot A_{\epsilon} \zeta \leq \frac{1}{\lambda_{\ominus}^0} A_{\epsilon} \zeta \cdot A_{\epsilon} \zeta \leq \frac{\lambda_{\oplus}^{\epsilon}}{\lambda_{\ominus}^0} A_{\epsilon} \zeta \cdot \zeta \leq \frac{\lambda_{\oplus}^{\epsilon}}{\lambda_{\ominus}^0} \lambda_2 A_0 \zeta \cdot \zeta$, we get a coarse estimate of the constant c_{\oplus} and conclude that

$$q(\rho_*) \leq \sqrt{1 - \frac{\lambda_1^2 \lambda_{\ominus}^0}{\lambda_2 \lambda_{\oplus}^{\epsilon}}} = \sqrt{1 - \left(\frac{\lambda_{\oplus}^{\epsilon} \lambda_{\ominus}^0}{\lambda_{\oplus}^{\epsilon} \lambda_{\oplus}^0} \right)^2} < 1. \quad (2.8)$$

However, in general the contraction factor can be better than in (2.8). Indeed, the value of ρ in (2.4) depends on the quantity $\max_{x \in \Omega} |\mathbb{B}^2(\rho, x, A_0, A_{\epsilon})|$, where

$$\mathbb{B}(\rho, x, A_0, A_{\epsilon}) := \mathbb{I} - \rho A_0^{-1}(x) A_{\epsilon}(x).$$

If A_0 is given, then the best $\rho_* > 0$ satisfies the condition

$$q(\rho_*, A_0, A_{\epsilon}) = \min_{\rho \in \mathbb{R}_+} \max_{x \in \Omega} |\mathbb{B}^2(\rho, x, A_0, A_{\epsilon})|. \quad (2.9)$$

In view of this fact, we obtain a criterium for selecting the best $A_0(x)$ among a certain collection \mathcal{A} of "simple" (e.g., polynomial or piece-wise constant) structures: find $\rho_* > 0$ and $A_0^* \in \mathcal{A}$ such that

$$q(\rho_*, A_0^*, A_{\epsilon}) := \max_{x \in \Omega} |\mathbb{B}^2(\rho_*, x, A_0^*, A_{\epsilon})| = \min_{A_0 \in \mathcal{A}} \min_{\rho \in \mathbb{R}_+} \max_{x \in \Omega} |\mathbb{B}^2(\rho, x, A_0, A_{\epsilon})|. \quad (2.10)$$

In the right hand side of (2.10) we have a matrix optimization problem. Solving it leads to the best simplified matrix A_0 and optimal ρ_* . It is worth outlining that this relatively simple problem does not require solutions of some auxiliary boundary value problems and can be done before iterative computations based on (2.6).

Particular case. Let $A_0 = a_0 \mathbb{I}$ and $A_\epsilon = a_\epsilon \mathbb{I}$. In this case, (2.8) means that ρ_* should be selected such that

$$\max_{x \in \Omega} |1 - \rho_* \mathfrak{h}(x)| = \min_{\rho} \max_{x \in \Omega} |1 - \rho \mathfrak{h}(x)|, \quad \mathfrak{h}(x) = \frac{a_\epsilon}{a_0} > 0. \quad (2.11)$$

If $d = 1$, then it is not difficult to show that

$$\max_x |1 - \rho \mathfrak{h}(x)| = \max\{|1 - \rho \underline{\mathfrak{h}}|, |1 - \rho \overline{\mathfrak{h}}|\},$$

where $\overline{\mathfrak{h}} := \max_{x \in \Omega} \mathfrak{h}(x)$ and $\underline{\mathfrak{h}} := \min_{x \in \Omega} \mathfrak{h}(x)$. We find that $\rho_* = \frac{2}{\underline{\mathfrak{h}} + \overline{\mathfrak{h}}}$ and

$$q(\rho_*, a_0, a_\epsilon) = \frac{\overline{\mathfrak{h}} - \underline{\mathfrak{h}}}{\underline{\mathfrak{h}} + \overline{\mathfrak{h}}} < 1.$$

For $a_0 = \text{const}$, we obtain $q(\rho_*, a_0, a_\epsilon) = \frac{\overline{a_\epsilon} - \underline{a_\epsilon}}{\overline{a_\epsilon} + \underline{a_\epsilon}}$.

If \mathcal{A} is a set of possible simplified coefficients, then the optimization problem (2.10) is reduced to finding $a_0^* \in \mathcal{A}$ that solves the problem

$$\min_{a_0 \in \mathcal{A}} \left(\frac{\max_{x \in \Omega} \frac{a_\epsilon}{a_0} - \min_{x \in \Omega} \frac{a_\epsilon}{a_0}}{\max_{x \in \Omega} \frac{a_\epsilon}{a_0} + \min_{x \in \Omega} \frac{a_\epsilon}{a_0}} \right). \quad (2.12)$$

Let a_ϵ be a function oscillating around a certain "mean" function a_0 . Assume that maximal relative deviation of a_0 from a_ϵ does not exceed δ , i.e., $\frac{a_\epsilon}{a_0} \in [1 - \mu_\ominus, 1 + \mu_\oplus]$, $\mu_\ominus \in (0, 1)$, $\mu_\oplus \geq 0$. The parameter $\delta := \mu_\ominus + \mu_\oplus$ characterises the scale of deviations. Since $\max_{x \in \Omega} \frac{a_\epsilon}{a_0} \leq 1 + \mu_\oplus$, and $\min_{x \in \Omega} \frac{a_\epsilon}{a_0} = 1 - \mu_\ominus$, we find that

$$q(\rho_*, a_0, a_\epsilon) = \frac{\delta}{2 + \mu_\oplus - \mu_\ominus} = 1 - \frac{1 - \mu_\ominus}{\delta/2 + 1 - \mu_\ominus}.$$

This formula shows that the method can be very efficient if δ is small, i.e., if a_ϵ oscillates around a_0 with a relatively small amplitude.

2.3 Discrete setting

Theorem (2.1) is applicable to the case where the problem is solved on a finite dimensional subspace $V_{0h} \in V_0$ associated with a mesh \mathcal{T}_h . Let $u_{0,h}$ solve the problem

$$\mathfrak{a}_0(u_{0,h}, w_h) = (f, w_h)_\Omega \quad \forall w_h \in V_{0h}$$

and the functions $\{u_{k+1,h}\} \in V_{0h}$, $k = 0, 1, 2, \dots$, satisfy

$$\mathfrak{a}_0(u_{k+1,h}, w_h) = \mathfrak{a}_0(u_{k,h}, w_h) - \rho \ell_{u_{k,h}}(w_h) \quad \forall w_h \in V_{0h}. \quad (2.13)$$

By repeating the same arguments as before, we conclude that $\{u_{k,h}\}$ tends to the fixed point $u_{\epsilon,h}$ of (2.13) provided that $\rho < 2\rho_*$. Obviously, $u_{\epsilon,h}$ satisfies the relation

$$0 = \ell_{u_{k,h}}(w_h) := a_\epsilon(u_{\epsilon,h}, \nabla w_h) - (f, w_h)_\Omega \quad \forall w_h \in V_{0h}. \quad (2.14)$$

We see that $u_{\epsilon,h}$ is the Galerkin approximation of u_ϵ on V_{0h} and obtain the a priori error estimate

$$\|u_{k,h} - u_{\epsilon,h}\|_0 \leq q^k(\rho) \|u_{0,h} - u_{\epsilon,h}\|, \quad (2.15)$$

which shows that approximations converge to $u_{\epsilon,h}$ with the geometric rate. From the practical point of view it is more important to have an estimate of $\|u_{k,h} - u_\epsilon\|_0$. In the next section we will obtain such estimates.

Now we discuss matrix equations that follow from (2.13) and compare them with the equations generated by a "straightforward" approach applied to (1.3). Let $\{\phi_i\}$, $i = 1, 2, \dots, N$ be a system of linearly independent trial functions and $V_h = \text{Span}\{\phi_i\}$. Define the vector $\mathbf{f} := \{f_i\}$, $f_i = (f, \phi_i)_\Omega$ and two matrixes $\mathbb{A}_\epsilon := \{a_\epsilon(\phi_i, \phi_j)\}$ and $\mathbb{A}_0 := \{a_0(\phi_i, \phi_j)\}$. In particular, we can use piecewise affine basis functions in $V_{0h} \subset H_0^1(\Omega)$ associated with the uniform tensor-product Cartesian grid. Denoting the fine grid size by $h = 1/(N+1)$, where N is the number of grid points in each spatial direction, the total problem size is estimated by N^d . For ease of exposition we assume that each scaled unit cell of univariate size $O(\epsilon)$ includes equal number n_0 of grid points.

Direct computation of $u_{\epsilon,h}$ requires solving the algebraic problem

$$\mathbb{A}_\epsilon \mathbf{v}_\epsilon = \mathbf{f}, \quad \mathbb{A}_\epsilon \in \mathbb{R}^{N^d \times N^d}, \quad \mathbf{f} \in \mathbb{R}^{N^d}, \quad (2.16)$$

with a sparse stiffness matrix \mathbb{A}_ϵ . Here $\mathbf{v}_\epsilon \in \mathbb{R}^{N^d}$ is the vector of nodal values that define $u_{\epsilon,h}$. The main bottleneck of the above computational scheme is due to the matrix size in the Galerkin system (2.16). Indeed, the univariate mesh parameter N is of the order of $N = O(\frac{n_0}{\epsilon})$, where n_0 is the mesh parameter that ensures the sufficient resolution of all data in the cell of length ϵ . In general, this parameter also depends on another structural parameter κ . Hence accurate approximations require huge values of N so that the numerical complexity of the direct solver for the system (2.16) scales polynomially in the frequency parameter $(1/\epsilon)^d$. Homogenization methods introduce a model simplification providing indirect $O(\sqrt{\epsilon})$ -approximation to the solution of (2.16). This method avoids inversion of \mathbb{A}_ϵ and leads to the problem (1.7), which can be solved on subspaces of much lower dimension.

The iteration scheme (2.13) suggests another way to avoid inversion of \mathbb{A}_ϵ . The basic iteration algorithm on the full finite element space (exact arithmetics) starts with $\mathbf{v}_0 = \mathbb{A}_0^{-1} \mathbf{f}$ and computes \mathbf{v}_{k+1} , $k = 0, 1, 2, \dots$ by solving the problem

$$\mathbb{A}_0 \mathbf{v}_{k+1} = \mathbb{A}_0 \mathbf{v}_k - \rho(\mathbb{A}_\epsilon \mathbf{v}_k - \mathbf{f}). \quad (2.17)$$

We can rewrite (2.17) in the form $\mathbf{v}_{k+1} - \mathbf{v}_k = \rho(\mathbb{A}_0^{-1} \mathbf{f} - \mathbb{A}_0^{-1} \mathbb{A}_\epsilon \mathbf{v}_k)$, which shows that (2.17) is equivalent to the iteration method applied to the preconditioned system

$$\mathbb{A}_0^{-1} \mathbb{A}_\epsilon \mathbf{u}_\epsilon = \mathbf{v}_0. \quad (2.18)$$

It is worth awaiting that in many cases (at least for periodic or almost periodic coefficients with small ϵ) the above selection of \mathbf{v}_0 with (\mathbb{A}_0) generated by the homogenized problem

(1.7) or another suitable simplified matrix) will provide a good starting approximation to the procedure (2.17). This fact was indeed confirmed in various numerical tests which show that such a constructed \mathbf{v}_0 is a good initial guess for the iteration method.

The iteration (2.17) involves only one operation with \mathbb{A}_ϵ : multiplication by the vector \mathbf{v}_k . If entries of the matrix \mathbb{A}_ϵ are generated by oscillating functions having low rank approximation in the so-called quantized tensor representations (QTT) [14], then (2.17) can be solved fairly easily by QTT-based tensor type methods applied to the properly transformed linear system (see Section 4). The approximate tensor arithmetics makes performing the iterations inexpensive. Thus, we obtain a new computational approach for a rather wide class of problems with periodic and quasi-periodic coefficients that allows to solve equation (2.16) by iteration (2.17) with the required precision at the cost that scales only logarithmically in ϵ .

3 Error control

For the control of approximation errors we use a posteriori estimates of the functional type (see [26, 27] and the references therein). They provide guaranteed and fully computable bounds of errors for any conforming approximation within the framework of a unified procedure, which does not require special features of approximations (e.g., exact satisfaction of the Galerkin orthogonality condition) or special features of the exact solution (e.g., extra regularity). Such estimates are robust and convenient for problems with complicated coefficients (see, e.g., [28]), where the above mentioned conditions are difficult to guaranty. We recall that for any approximation $v \in V$ of the problem (1.1) we have the following estimate of the error $e := u_\epsilon - v$

$$\|e\|_\epsilon \leq \|A_\epsilon \nabla v - y\|_\epsilon + C_\Omega \|\operatorname{div} y + f\| := M_\oplus(v, y), \quad (3.1)$$

where $\|e\|_\epsilon^2 = a_\epsilon(e, e)$, y is an arbitrary vector function in $H(\Omega, \operatorname{div})$ and C_Ω is the Friedrichs constant. If $\Omega \subset \{x \in \mathbb{R}^d \mid a_s < x < b_s, b_s - a_s = l_s, s = 1, 2, \dots, d\}$, then $C_\Omega = \frac{1}{\kappa\pi}$, $\kappa^2 = \sum_{s=1}^d \frac{1}{l_s^2}$. In particular, $C_\Omega = \frac{1}{d^{1/2}\pi}$ for $\Omega = (0, 1)^d$.

The functional $M_\oplus(v, y) : H_0^1(\Omega) \times H(\Omega, \operatorname{div}) \rightarrow \mathbb{R}_+$ is an error majorant. Properties of such error majorants are well studied (see, e.g., [27]). We know that $M_\oplus(v, y)$ vanishes if and only if $v = u_\epsilon$ and $y = p_\epsilon := A_\epsilon \nabla u_\epsilon$. Moreover, for any $v \in H_0^1(\Omega)$ the functional $M_\oplus(v, p_\epsilon)$ coincides with the error and the integrand of $M_\oplus(v, p_\epsilon)$ shows the distribution of local errors. Numerous tests performed for different boundary value problems have confirmed practical efficiency of this and other error majorants derived for various problems. It was shown that M_\oplus is a guaranteed and efficient majorant of the global error and good indicator of local errors if the exact flux is replaced by a certain numerical reconstruction p_h (in our case instead of p_ϵ we use $p_{h,\epsilon}$). There are many different ways to obtain suitable reconstructions (e.g., see [22] for a systematic discussion of computational aspects of this error estimation method).

Error majorants can be efficiently used for the evaluation of modeling errors (see [30, 29]). In particular, if we set $y = A_0 \nabla u_0$, then (3.1) implies a simple estimate of the modeling error caused by simplification of coefficients: $\|e\|_\epsilon \leq \|(A_\epsilon - A_0) \nabla v\|_\epsilon$.

However, for problems with highly oscillating coefficients the general majorant (3.1) has a substantial drawback: it contains a norm generated by A_ϵ^{-1} . In our analysis we try to avoid all the operations related to this most complicated matrix except multiplication by a vector (which can be performed by tensor type methods). Hence, the goal is to modify general a posteriori estimates in accordance with this principle. Consider one step of the iteration method, where the function $v \in V_0$ generates $u = Tv \in V_0$. For a contractive mapping T we have two sided error estimates (see [25, 27, 32]), which imply

$$\|v - u_\epsilon\|_0 \in \left\{ \frac{1}{1 + q(\rho)}, \frac{1}{1 - q(\rho)} \right\} \|u - v\|_0 \quad (3.2)$$

Here v is known, but $u = Tv$ is generally unknown and we need to use some approximation \tilde{u} instead. It is easy to see that

$$\|u - v\|_0 \leq \|\eta\|_0 + \|u - \tilde{u}\|_0, \quad (3.3)$$

$$\|u - v\|_0 \geq \|\eta\|_0 - \|u - \tilde{u}\|_0, \quad (3.4)$$

where the function $\eta := \tilde{u} - v$ is known. The estimates (3.3) and (3.4) would be fully computable if we find a computable majorant of the norm $\|\tilde{u} - u\|_0$.

We note that

$$a_0(u, w) = a_0(v, w) - \rho \int_{\Omega} (A_\epsilon \nabla v \cdot \nabla w - fw) \quad \forall w \in V_0. \quad (3.5)$$

Then for any $y \in H(\Omega, \text{div})$ and $w \in V_0$, we have

$$\begin{aligned} a_0(u - \tilde{u}, w) &= a_0(v - \tilde{u}, w) - \rho \int_{\Omega} (A_\epsilon \nabla v \cdot \nabla w - fw) \\ &= \int_{\Omega} (A_0 \nabla(v - \tilde{u}) \cdot \nabla w - \rho(A_\epsilon \nabla v \cdot \nabla w - fw)) dx \\ &= \int_{\Omega} (A_0 \nabla(v - \tilde{u}) - \rho A_\epsilon \nabla v + y) \cdot \nabla w + (\rho f + \text{div} y) w dx. \end{aligned} \quad (3.6)$$

Using the notation $\tau := y - \rho A_\epsilon \nabla v$, we obtain

$$\begin{aligned} \|u - \tilde{u}\|_0 &\leq \left(\int_{\Omega} (A_0 \nabla \eta \cdot \nabla \eta + A_0^{-1} \tau \cdot \tau - 2 \nabla \eta \cdot \tau) dx \right)^{1/2} \\ &\quad + \frac{C_F}{\lambda_{\ominus}^0} \|\text{div} y + \rho f\| =: \mathcal{M}_{\oplus}(\tilde{u}, v, y). \end{aligned} \quad (3.7)$$

Note that

$$\inf_{y \in H(\Omega, \text{div})} \mathcal{M}_{\oplus}(\tilde{u}, v, y) = \|u - \tilde{u}\|_0.$$

Indeed, set $y = A_0 \nabla(u - v) + \rho A_\epsilon \nabla v$. Then, $\tau = A_0 \nabla(u - v)$, $\operatorname{div} y + \rho f = 0$, and the first term of the majorant is equal to $\|u - \tilde{u}\|_0$. Hence, the estimate has no gap.

Now we apply these relations to the step k of (2.6). Set $v = u_{k,h}$ (approximation computed at step k using a mesh \mathcal{T}_h^k). Then $u = Tu_{k,h}$ is the exact solution of (3.5), which we do not know. Instead we have a function $\tilde{u} = u_{k+1,h}$ computed on the mesh \mathcal{T}_h^{k+1} (it may coincide with the previous mesh \mathcal{T}_h^k or be a new one constructed by, e.g., a refinement procedure). The function $\eta = \eta_{k+1} := u_{k+1,h} - u_{k,h}$ is known. By (3.3) and (3.4) we obtain

$$\|u_{k,h} - u_\epsilon\|_0 \leq \frac{1}{1-q} (\|\eta_{k+1}\|_0 + \mathcal{M}(u_{k+1,h}, u_{k,h}, y)), \quad (3.8)$$

$$\|u_{k,h} - u_\epsilon\|_0 \geq \frac{1}{1+q} (\|\eta_{k+1}\|_0 - \mathcal{M}(u_{k+1,h}, u_{k,h}, y)). \quad (3.9)$$

Here y is any vector function in $H(\Omega, \operatorname{div})$. Certainly, getting minimal values of the majorant require a suitable numerical reconstruction of the exact flux of the problem (2.6), which is $q_k = A_0 \nabla u_{k+1} - \sigma_k$, where $\sigma_k = (A_0 - \rho A_\epsilon) \nabla u_{k,h}$ is known. Since the coefficients of A_0 are regular and do not oscillate, reconstructions of such a flux can be done by well known methods (see, e.g., [22, 27] and the literature cited in these books). We do not discuss this question in detail because it will be the matter of a special publication focused on multidimensional problems.

If $d = 1$ (related to the numerical tests below), then the flux is easy to reconstruct. In this case the problem (1.1) is $(a_\epsilon u'_\epsilon)' + f = 0$ and the simplified problem is $(a_0 u'_0)' + f = 0$. We set $y = \rho(-g(x) + c)$ (where $g(x) = \int_0^x f dx$) and define the constant c by minimizing the first term of $\mathcal{M}(u_{k+1,h}, u_{k,h}, y)$. We have $\tau = \rho(c - g(x) - a_\epsilon u'_{k,h})$ and need to find c minimizing the quantity

$$\int_0^1 (a_0 |\eta'_{k+1}|^2 + a_0^{-1} \rho^2 (c - g(x) - a_\epsilon u'_{k,h})^2 - 2\rho \eta'_{k+1} (c - g(x) - a_\epsilon u'_{k,h})) dx.$$

Since $\int_0^1 \eta'_{k+1} dx = 0$, the problem is reduced to minimization of the second term and the best c satisfies the equation $\int_0^1 a_0^{-1} (c - g(x) - a_\epsilon u'_{k,h}) dx = 0$. Hence $c = c_k := \left(\int_0^1 a_0^{-1} (g(x) + a_\epsilon u'_{k,h}) dx \right) \left(\int_0^1 a_0^{-1} dx \right)^{-1}$ and

$$\begin{aligned} \mathcal{M}^2(u_{k+1,h}, u_{k,h}, y) &= \\ &= \int_0^1 (a_0 |\eta'_{k+1}|^2 + a_0^{-1} \rho^2 (c_k - g(x) - a_\epsilon u'_{k,h})^2 + 2\rho \eta'_{k+1} (g(x) + a_\epsilon u'_{k,h})) dx. \end{aligned}$$

We see that the majorant is fully computable. Moreover, if the sequence $\{u_{k,h}\}$ converges to u_ϵ in V , then $\|\eta_{k+1}\|_0 \rightarrow 0$. Hence the first and the last terms of the above integral tend to

zero. Also, $g(x) + a_\epsilon u'_{k,h} \rightarrow g(x) + a_\epsilon u'_\epsilon =: c_*$ in L^2 . Note that

$$c_k = \frac{\int_0^1 a_0^{-1}(g(x) + a_\epsilon u'_{k,h})dx}{\int_0^1 a_0^{-1} dx} = c_* + \frac{\int_0^1 a_0^{-1} a_\epsilon (u'_{k,h} - u'_\epsilon)dx}{\int_0^1 a_0^{-1} dx}$$

Therefore, $|c_k - c_*| \leq \mu_{0,\epsilon} \|u'_{k,h} - u'_\epsilon\|_0$, where $\mu_{0,\epsilon} = \left(\int_0^1 a_0^{-1} a_\epsilon^2 dx \right)^{1/2} \left(\int_0^1 a_0^{-1} dx \right)^{-1}$. does not depend on k . Therefore, the second term also tends to zero and we conclude that the majorant (3.8) tends to zero. Also, it is possible to show that the majorant is equivalent to the error.

4 Tensor-based preconditioned iterative scheme

The main concept of our tensor-based approach is the direct iterative solution of the initial large algebraic system (2.16) in the form of preconditioned iteration (2.17) by using low-parametric data formats, exploiting certain redundancy in the grid-based representation of matrices and vectors involved. This is realized, first, by transformation of the “low-dimensional” FEM-Galerkin equation to the equivalent system posed in the high dimensional quantized tensor space, and then by solving this system iteratively using the low-rank QTT tensor approximation [14] to the Galerkin stiffness matrix, the preconditioner and all vectors involved. This allows to compute the numerical approximation to the exact solution discretized on a fine grid up to the chosen precision $\delta > 0$, adapted to the mesh-resolution but independent on the frequency parameter $1/\epsilon$.

The approach is well adapted to fast QTT-based tensor approximation method, what is natural to await because the QTT tensors fit well the intrinsic features of FEM discretizations to functions and operators generated via periodic and quasi-periodic geometric structures [14, 11, 13, 16]. The numerical cost of the rank-structured iteration can be bounded by $O(|\log \epsilon|^q)$ provided that rank parameters remain small.

4.1 QTT tensor representation of function related vectors and matrices

In this section we present a brief overview of QTT tensor approximation method [14] which is the base for the construction of the presented tensor-based computational scheme. We refer to surveys on commonly used low-rank representations of discrete functions and operators [18, 15, 31].

The QTT-type approximation of an N -vector with $N = q^L$, $L \in \mathbb{N}$ (usually $q = 2$) allows to reduce the asymptotic storage cost to $O(\log N)$ [14]. The QTT rank decomposition applies to a tensor obtained by the q -adic folding (reshaping) of the target long vector to an L -dimensional $q \times \dots \times q$ data array considered in the L -dimensional quantized tensor space. As the basic result, in [14] it was shown that for a large class of function related vectors (tensors) such a procedure allows the low-rank representation of their quantized

L -dimensional image, thus reducing the representation complexity to the logarithmic scale $O(\log N)$.

In particular, a vector $\mathbf{x} = [x(i)]_{i=1}^N \in \mathbb{R}^N$, is reshaped to its quantics image in $\mathbb{Q}_{q,L} = \bigotimes_{j=1}^L \mathbb{K}^q$, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, by q -adic folding,

$$\mathcal{F}_{q,L} : \mathbf{x} \rightarrow \mathbf{Y} = [y(\mathbf{j})] \in \mathbb{Q}_{q,L}, \quad \mathbf{j} = \{j_1, \dots, j_L\}, \quad \text{with } j_\nu \in \{1, 2, \dots, q\},$$

where for fixed i , we have $y(\mathbf{j}) := x(i)$, and $j_\nu = j_\nu(i)$, $\nu = 1, \dots, L$, is defined via q -coding, $j_\nu - 1 = C_{-1+\nu}$, such that the coefficients $C_{-1+\nu}$ are found from the q -adic representation of $i - 1$,

$$i - 1 = C_0 + C_1 q^1 + \dots + C_{L-1} q^{L-1} \equiv \sum_{\nu=1}^L (j_\nu - 1) q^{\nu-1}.$$

Suppose that the quantized image for certain N -vector (i.e. an element of L -dimensional quantized tensor space $\mathbb{Q}_{q,L}$ with $L = \log_q N$) can be effectively represented (approximated) in the low-rank canonical or TT format. For given QTT-rank parameters $\{r_k\}$ ($k = 1, \dots, L - 1$) the number of representation parameters in the QTT approximation can be estimated by

$$qr^2 \log_q N \ll N, \quad \text{where } r_k \leq r, \quad k = 1, \dots, L - 1,$$

providing log-volume scaling in the size of initial vector, N . The optimal choice of the base q is shown to be $q = 2$ or $q = 3$ [14], however the numerical realizations are usually implemented by using binary coding, i.e. for $q = 2$. For $d \geq 2$ the construction is similar [14].

The favorable features of the QTT approximation method are due to the perfect low rank decompositions discovered for the wide class of function-related tensors [14]:

Proposition 4.1 ([14]) *Let vector $\mathbf{x} \in \mathbb{C}^N$, $N = 2^L$, be obtained by sampling a continuous function $f \in C[0, 1]$ over the uniform grid of size N . Then the following QTT-rank estimates are valid independent on the vector size N :*

- (A) $r = 1$ for complex exponentials, $f(x) = e^{i\omega x}$, $\omega \in \mathbb{R}$.
- (B) $r = 2$ for trigonometric functions, $f(x) = \sin \omega x$, $f(x) = \cos \omega x$, $\omega \in \mathbb{R}$.
- (C) $r \leq m + 1$ for polynomials of degree m .
- (D) For a function f with the QTT-rank r_0 modulated by another function g with the QTT-rank r (say, step-type function, plain wave, polynomial) the QTT rank of a product fg is bounded by a multiple of r and r_0 .
- (E) QTT rank for the periodic amplification of a reference function on a unit cell to a rectangular lattice is of the same order as that for the reference function [11].

Concerning the matrix case, it was found in [24] by numerical tests that in some cases the dyadic reshaping of an $N \times N$ matrix with $N = 2^L$ may lead to a small TT-rank of the resultant matrix rearranged to the tensor form. The explicit low-rank QTT representations for a class of discrete multidimensional matrices mapping the space $\mathbb{Q}_{2,L}$ into itself were proven in [8], see also [15].

In our applications the concept for the construction of fast numerical methods is based on the ϵ -independent low-rank QTT approximation of all function-related vectors and operator-related matrices involved in the computational scheme. This allows to reduce the numerical

costs to the logarithmic scale in the grid-size, i.e. of order $O(|\log \epsilon|^q)$. The critical issue concerns with QTT approximation of the FEM-Galerkin matrix \mathbb{A}_ϵ generated by the highly oscillating or/and jumping diffusion coefficient, and of the respective “homogenized” preconditioner \mathbb{A}_0 .

4.2 The Galerkin FEM scheme

For further discussion we choose the Galerkin FEM with N piecewise-linear hat functions $\{\phi_i\}$ in the physical domain $\Omega = [0, 1]$, constructed on a fine uniform grid with step size $h = 1/(N + 1)$, which is a small fraction of ϵ , and nodes $x_i = hi$, $i = 1, \dots, N$. For ease of notation we further denote $a_\epsilon(x) = A_\epsilon(x)$, then the entries of the exact stiffness matrix $\mathbb{A}[a_\epsilon]$ read

$$(\mathbb{A}[a_\epsilon])_{i,i'} = (a_\epsilon(x) \nabla \phi_i(x), \nabla \phi_{i'}(x))_{L_2(D)}, \quad i, i' = 1, \dots, N. \quad (4.1)$$

To simplify the approximation procedure, we may assume that the coefficient remains constant at each spatial interval $[x_{i-1}, x_i]$, which corresponds to the evaluation of the scalar product above via the midpoint quadrature rule. It is known that this quadrature yields the approximation order $\mathcal{O}(h^2)$, the same as the piecewise-linear discretization of the solution.

We introduce the coefficient vector $\mathbf{a} = [a_i] \in \mathbb{R}^N$, $a_i = a_\epsilon(x_{i-1/2})$, $i = 1, \dots, N$, then the resulting tridiagonal matrix takes the form,

$$\mathbb{A}[\mathbf{a}] = \frac{1}{h} \begin{bmatrix} a_1 + a_2 & -a_2 & & & \\ -a_2 & a_2 + a_3 & -a_3 & & \\ & \ddots & \ddots & \ddots & \\ & & -a_{N-1} & a_{N-1} + a_N & -a_N \\ & & & -a_N & 2a_N \end{bmatrix}. \quad (4.2)$$

4.3 QTT tensor representation of the system matrix

In this section we discuss the low-rank QTT representations of the Galerkin matrices approximating elliptic operators with variable coefficients in 1D. In particular, we construct the QTT representation of arising three-diagonal stiffness matrices by using the related results in [10]. The approach can be extended to d -dimensional equations defined on the lattice-type geometries.

We let $N = 2^L$ and represent the indices in physical space in the binary coding

$$i = \overline{i_1, \dots, i_L}, \quad \text{where } L = \log_2 N,$$

to consider the QTT decomposition of vectors and matrices involved in the discrete problem. Then the coefficient vector $\mathbf{a} = [a_i]$ in (4.2) can be represented in the rank- \mathbf{r} QTT form as an L -dimensional tensor, where $\mathbf{r} = (r_1, \dots, r_{L-1})$:

$$a_i = \sum_{k_1, \dots, k_{L-1}=1}^{r_1, \dots, r_{L-1}} a_{k_1}^{(1)}(i_1) \cdots a_{k_{L-1}}^{(L)}(i_L). \quad (4.3)$$

The QTT cores of the matrix $\mathbb{A}[\mathbf{a}]$ can be written similarly, recalling that the vector is turned to the diagonal matrix without changing the TT ranks. Specifically, the matrix $\mathbb{A}[\mathbf{a}]$

can be brought into the QTT format by using the shift matrices and their explicit rank-2 QTT representation. Let us denote by $\mathbb{S} = [s_{i,i'}] \in \mathbb{R}^{N \times N}$ the upper shift matrix given by

$$s_{i,i'} = \begin{cases} 1, & i' = i + 1, \\ 0, & \text{else,} \end{cases},$$

and notice that this matrix has exact rank-2 QTT representation [8]. Then it holds

$$\mathbb{A}[\mathbf{a}] = \mathbb{S} \text{diag}(\mathbf{a}) + \text{diag}(\mathbf{a} + \mathbb{S}\mathbf{a}) + \text{diag}(\mathbf{a})\mathbb{S}^\top, \quad (4.4)$$

with the maximal QTT ranks estimate $r(\mathbb{A}[\mathbf{a}]) \leq 7r(\mathbf{a})$, controlled by the QTT rank of the coefficients vector \mathbf{a} . The representativity of the stiffness matrix in low-rank formats can be summarized as follows [10].

Theorem 4.2 *Let the problem be discretized by the matrix (4.2) with the use of the Galerkin-FEM method on the uniform grid. Suppose that the diffusion coefficient vector \mathbf{a} is given in a form of a QTT decomposition (4.3) with the following QTT-rank bounds $r_p \leq R$, for $p = 1, \dots, L - 1$. Then the QTT ranks of the matrix $\mathbb{A}[\mathbf{a}]$ can be bounded by $7R$.*

Theorem 4.2 ensures that the arising Galerkin system of linear equations (2.16) can be efficiently solved iteratively as (2.17) by using the low-rank QTT representation of each of entities \mathbf{a} , \mathbf{u}_ϵ , and \mathbf{f} , provided that the spectrally close preconditioner, \mathbb{A}_0 , to the stiffness matrix $\mathbb{A}_\epsilon[\mathbf{a}]$ is constructed.

4.4 Main assumptions and the construction of preconditioner

In the following numerical tests we chose the homogenized elliptic operator defined by the coefficient $a_0(x)$ as the preconditioner.

In the case of exotic coefficients, the construction of “homogenized” coefficient may depend on the shape of the initial equation coefficient $a_\epsilon(x)$. We define the preconditioning operator \mathbb{A}_0 via the generalized averaging procedure

$$\tilde{a}_0(x) = \frac{1}{2}(a^+(x) + a^-(x)),$$

where $a^+(x)$ and $a^-(x)$ are chosen as *majorants and minorants* of $a_\epsilon(x)$, respectively. Examples of such a construction is given in Figure 1.1, left and middle, and in Figure 5.1, where the coefficient $\tilde{a}_0(x)$ is colored in red. The estimate on the condition number of the preconditioned matrix is given by the following simple lemma.

Lemma 4.3 *Define $q(x) := (a^+(x) - \tilde{a}_0(x))/\tilde{a}_0(x)$, then the condition number of the preconditioned matrix $\mathbb{A}_0^{-1}\mathbb{A}_\epsilon$ is bounded by*

$$\text{cond}\{\mathbb{A}_0^{-1}\mathbb{A}_\epsilon\} \leq C \max \frac{1 + q(x)}{1 - q(x)}.$$

The main assumptions for applicability of the presented tensor method are the following:

(A) Sampling vectors for *homogenized* coefficient $\tilde{a}_0(x)$, for the oscillatory one $a_\epsilon(x)$, as well as for the loading vector \mathbf{f} all have low QTT ranks.

(B) Numerical implementation of \mathbb{A}_0^{-1} is cheap, the solution of “homogenized” equation $\mathbb{A}_0 \mathbf{v}_0 = \mathbf{f}$ has low QTT rank.

(C) For FEM-Galerkin approximation matrix \mathbb{A}_ϵ the spectral equivalence relation holds

$$\lambda_0 \mathbb{A}_0 \leq \mathbb{A}_\epsilon \leq \lambda_1 \mathbb{A}_0, \quad \lambda_0, \lambda_1 > 0.$$

Assumptions (A) – (C) are satisfied in all numerical examples presented in the following.

The tensor iterative scheme with QTT rank truncation is described as follows. Choose the threshold parameter $\delta > 0$, and denote by \mathcal{T}_δ the tensor operation producing almost the best QTT δ -approximation to the target rank-structured tensor. Then the exact iteration (2.17) on the full finite element space is modified as follows: Starts with $\mathbf{v}_0 = \mathbb{A}_0^{-1} \mathbf{f}$ presented as the low QTT rank tensor and then compute \mathbf{v}_{k+1} , $k = 0, 1, 2, \dots$ via fix-point iteration performed in the quantized tensor space $\mathbb{Q}_{2,L}$ and accomplished with δ -rank truncation,

$$\mathbf{v}_{k+1} = \mathcal{T}_\delta (\beta \mathbf{v}_0 - \mathbb{B} \mathbf{v}_k), \quad \text{with} \quad \mathbb{B} = \beta \mathbb{A}_0^{-1} \mathbb{A}_\epsilon - \mathbb{E}. \quad (4.5)$$

Condition (C) ensures the geometric convergence rate for the PCG, Preconditioned Steepest Descent (PSD) and others accelerated preconditioned iterations performed in the tensor format (4.5).

Taking into account assumptions (A) – (C), we arrive at the efficient preconditioned iterative solver for the initial FEM system of equations (2.16) discretized on finest grid of size h that resolves all local peculiarities in the matrix coefficients, i.e. $\epsilon \approx n_0 h$. The natural choice of the rank truncation parameter might be $\delta = O(h^2)$.

Summary 4.4 *Our model reduction approach introduces ϵ -adapted tensor structured approximation to the initial PDE and to the corresponding preconditioner (in turn, based on certain averaging of the oscillating coefficient) that has low-parametric representation as the QTT tensor and fits well the almost periodic structure in the coefficients and in the solution, uniformly in the frequency parameter $1/\epsilon$. Under certain assumptions on the quality of the quantized tensor approximation to the input data and the solution, the numerical complexity can be reduced to the logarithmic scale, $O(|\log \epsilon|^q)$.*

5 Numerics: iterative solver with logarithmic complexity

5.1 Description of problem classes

We consider several classes of oscillating or/and jumping coefficients. Our example of the ideal periodic problem is described by the family of diffusion coefficients

$$A_\epsilon(x) = C + \sin(\omega x) > 0, \quad x \in \Omega, \quad (5.1)$$

where the frequency $\omega \in \mathbb{R}$ (i.e. $\epsilon = 1/\omega$) may be chosen as an arbitrarily large constant, see Fig. 5.1, left. In this case the rank-2 QTT representation of the coefficients vector is suggested in [14], see Proposition 4.1, (B).

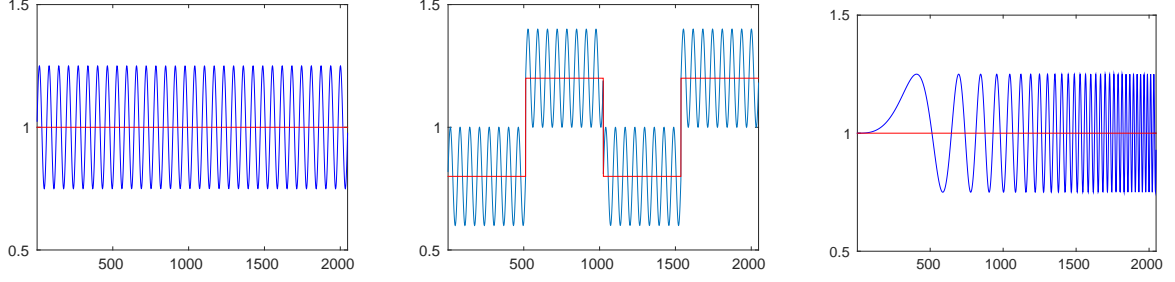


Figure 5.1: Examples of periodic and non-periodic oscillating coefficients.

The second class of “modulated periodic” coefficients is defined by

$$A_\epsilon(x) = C + g(x) \sin(\omega x), \quad (5.2)$$

where the modulating function $g(x) > 0$ should be chosen in such a way that the rank of the QTT approximation to the corresponding function related vector representation of $g(x)$ remains small. In case of modulated periodic coefficients the QTT rank of the modulated function $A_\epsilon(x)$ is bounded by the product of QTT-ranks for the modulator and oscillator, that is the well known property of Hadamard product of tensors, see Proposition 4.1, (D). For the particular choice of the modulator $g(x)$ given by a multi-step function, see Fig. 5.1, middle, the QTT-ranks are exactly 2. In this case Theorem 4.2 ensures that the QTT rank of the resultant equation coefficient \mathbf{a} does not exceed 4.

The third class is described by “exotic” oscillators which may have nonlinear highly oscillating behavior and could not be treated by the conventional homogenization methods. The coefficient is given by

$$A_\epsilon(x) = C + g(x) \sin(\omega x^m), \quad m = 2, 3, \dots, \quad (5.3)$$

see Fig. 5.1, right, where $C = 1$, $g(x) = 1$, and $m = 3$. Clearly, the first two classes of coefficients are the particular cases of “exotic” oscillator in (5.3).

In the general case of “exotic” oscillators (5.3) the explicit QTT-rank bounds are not known, however, in most examples considered so far the numerical tests indicate (see e.g. Table 5.1) the low-rank QTT approximations with high accuracy (numerical justification). The rigorous QTT approximation analysis is possible for some special classes of oscillating functions, see [16].

5.2 Numerics and comments

In the following numerical tests we use the simple preconditioner matrix corresponding to the constant coefficient a_0 defined by the mean value of the initial highly-oscillating function a_ϵ , i.e. $a_0 = \{a_\epsilon\}$. In this case \mathbb{A}_0 is just the scaled 1D Laplacian.

We apply the Preconditioned Steepest Descent (PSD) iteration with the QTT-rank truncation up to given $\delta > 0$ to solve the preconditioned system of linear equations represented in the high-dimensional quantized tensor space. In all the experiments we specify the rank truncation tolerance by $\delta = 10^{-7}$. The frequency $\omega = 1/\epsilon$ was parametrized in the form

$N = 2^L$, iter.	2^{13} , (it)	2^{14} , (it)	2^{15} , (it)	2^{16} , (it)	2^{17} , (it)	$r(\mathbf{a}_\epsilon)$	$r(\mathbf{u}_\epsilon)$
4-steps coef.	3.4, (9)	4.3, (9)	4.5, (9)	6.7, (9)	14.3, (14)	2.9	4.96
$C + \sin(\omega x)$	0.97, (5)	1.2, (5)	1.3, (5)	2.0, (6)	2.1, (6)	2.67	3.7
$C + \sin(\omega x^3)$	5.3, (5)	10.0, (6)	9.95, (6)	11.98, (6)	16.2, (5)	7.53	8.24

Table 5.1: CPU times (sec) and rank bounds for the three types of oscillating coefficients.

$\omega = 2\pi K$, with the particular choice $K = 64$. The numerical results with smaller or larger frequency parameter K demonstrated the similar features.

We observe the uniform geometric convergence of PSD iteration for each of three examples considered above, see Table 5.1 and Fig. 5.2. Moreover, results in Table 5.1 indicate the only logarithmic growth of CPU time per iteration with respect to the grid-size N . A posteriori error control indicates the convergence up to 10^{-5} in the H_1 -norm on the finest grid.

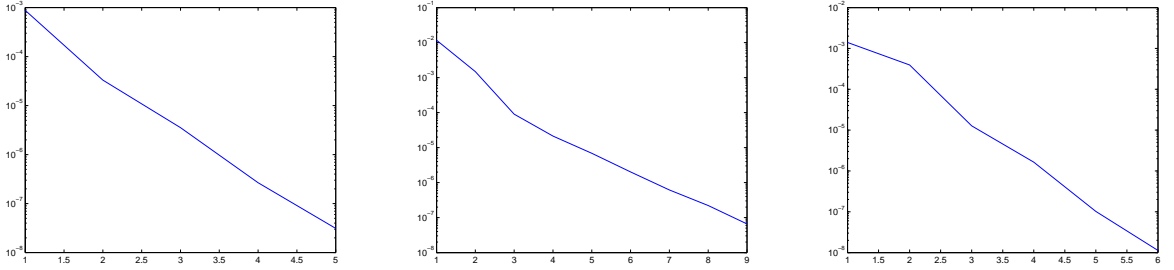


Figure 5.2: PSD iteration history for the periodic (left), jumping oscillating, and cubically oscillating (right) coefficients, see Fig. 5.1.

The exact solution \mathbf{u}_ϵ is calculated with the approximation error of the same order δ (independently of ϵ) that improves dramatically the approximation error of the homogenized solution. Notice that in all cases presented in Table 5.1 the following numerical precision has been achieved $\|u_\epsilon - \mathbf{v}_\delta\|_0 \simeq 10^{-7}$, and $\|u_\epsilon - \mathbf{v}_\delta\|_1 \simeq 10^{-6}$.

Though the difference between exact and homogenized solutions may be of order $O(\sqrt{\epsilon})$, see Fig. 5.3, the residual remains to be large as indicated in Fig. 5.4, no convergence in higher derivatives as $\epsilon \rightarrow 0$ is observed.

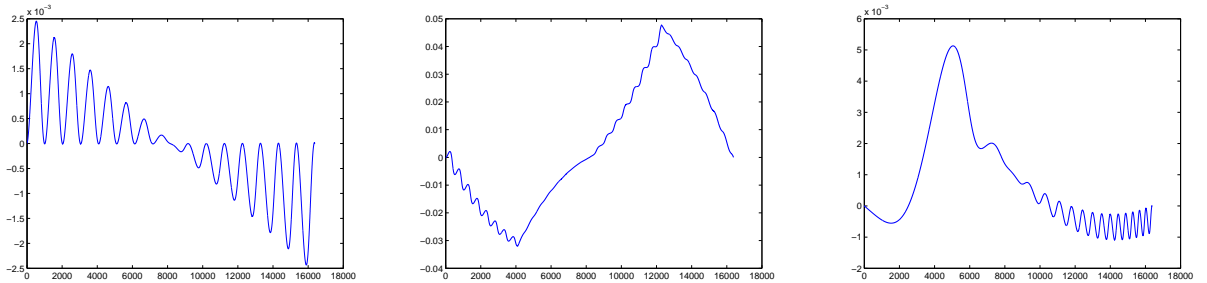


Figure 5.3: Difference between exact and homogenized solutions for three cases in Fig. 5.1.

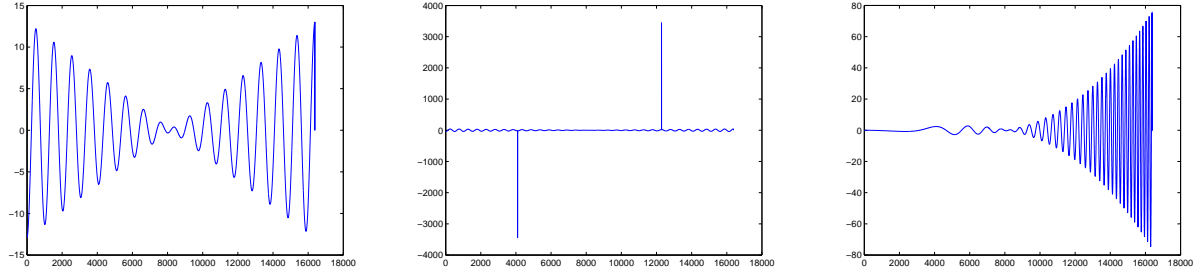


Figure 5.4: Residual for the homogenized periodic (left), 4-step jumping, and cubically oscillating (right) solutions.

In the case modulated periodic coefficients (see Fig. 5.1, middle, right) the standard homogenization theory does not provide the convergence even in the limit of small parameter ϵ . In this case Figure 5.5 demonstrates the systematic error between the exact (red) and homogenized solutions.

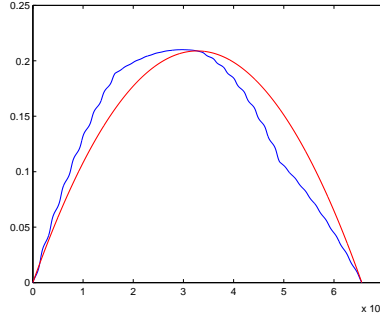


Figure 5.5: Exact (red) and homogenized solutions (4-jumping coefficients in Fig. 5.1, middle).

Acknowledgment. The second author thanks Max Planck Institute for Mathematics in the Sciences in Leipzig for support.

References

- [1] Bakhvalov, N. S., Panasenko, G. Homogenisation: Averaging Processes In Periodic Media: Mathematical Problems In The Mechanics Of Composite Materials. Springer, 1989.
- [2] Bensoussan, A., Lions, J.-L., Papanicolaou, G. (1978): Asymptotic analysis for periodic structures. Amsterdam: North-Holland
- [3] Cioranescu, D., Donato, P. (1999): An Introduction to Homogenization. Oxford Lecture Series in Mathematics and its Applications. Bd. 17. Oxford University Press
- [4] Antoine Gloria and Felix Otto. *An optimal error estimate in stochastic homogenization of discrete elliptic equations*. In: The annals of applied probability, 22 (2012) 1, p. 1-28.
- [5] Jikov, V.V., Kozlov, S.M., Oleinik, O.A. (1994): Homogenization of differential operators and integral functionals. Berlin: Springer
- [6] Friedman, A. (1976): Partial Differential Equations. R. E. Krieger Pub. Co., Huntington, NY

- [7] R. Glowinski, J.-L. Lions, R. Trémolierés. *Analyse numérique des inéquations variationnelles*. Dunod, Paris, 1976.
- [8] V. Kazeev, and B.N. Khoromskij. *Explicit low-rank QTT representation of Laplace operator and its inverse*. SIAM Journal on Matrix Anal. and Appl., 33(3), 2012, 742-758.
- [9] S.V. Dolgov, B.N. Khoromskij, I. Oseledets, and E.E. Tyrtysnikov. *Low-rank tensor structure of solutions to elliptic problems with jumping coefficients*. J. of Comput. Math. v. 30, No. 1, 2012, 14-23.
- [10] S. Dolgov, V. Kazeev, and B.N. Khoromskij. *The tensor-structured solution of one-dimensional elliptic differential equations with high-dimensional parameters*. Preprint 51/2012, MPI MiS, Leipzig 2012 (submitted).
- [11] V. Khoromskaia and B. N. Khoromskij. *Grid-based lattice summation of electrostatic potentials by assembled rank-structured tensor approximation*. Comp. Phys. Communications, 185 (2014), pp. 3162-3174. DOI: 10.1016/j.cpc.2014.08.015.
- [12] V. Khoromskaia, and B.N. Khoromskij. *Tensor Approach to Linearized Hartree-Fock Equation for Lattice-type and Periodic Systems*. Preprint 62/2014, MPI MiS, Leipzig 2014 (submitted). E-preprint arXiv:1408.3839, 2014.
- [13] V. Khoromskaia and B. N. Khoromskij. *Assembled Tucker tensor method to grid-based summation of long-range potentials on 3D lattices with defects*. Preprint 88/2014, MPI MiS, Leipzig 2014 (submitted). E-preprint arXiv:1411.1994, 2014.
- [14] B.N. Khoromskij. *$O(d \log N)$ -Quantics Approximation of N -d Tensors in High-Dimensional Numerical Modeling*. Constr. Approx. 34 (2011) 257–280.
- [15] B.N. Khoromskij. *Tensors-structured Numerical Methods in Scientific Computing: Survey on Recent Advances*. Chemometr. Intell. Lab. Syst. 110 (2012), 1-19.
- [16] Boris N. Khoromskij, and A. Veit. *Efficient computation of highly oscillatory integrals by using QTT tensor approximation*. E-preprint arXiv:1408.5224, 2014 (submitted).
- [17] T. Kailath, and A. Sayed. *Fast reliable algorithms for matrices with structure*. SIAM Publication, Philadelphia, 1999.
- [18] T. G. Kolda and B. W. Bader. *Tensor Decompositions and Applications*. SIAM Rev. 51(3) (2009) 455–500.
- [19] Yu. A. Kuznetsov. Two-level preconditioners with projectors for unstructured grids, *Russ. J. of Numer. Anal. Math. Modelling*, 15(2000), No. 3-4, 247–255.
- [20] Yu. Kuznetsov. Mixed FE method with piece-wise constant fluxes on polyhedral meshes, *Russ. J. of Numer. Anal. Math. Modelling*, 29(2014), No. 4, 231–238.
- [21] Yu. Kuznetsov and S. Repin. New mixed finite element method on polygonal and polyhedral meshes, *Russ. J. Numer. Anal. Math. Modelling*, Vol. 18(2003), 261–278.
- [22] O. Mali, P. Neittaanmaki, S. Repin. *Accuracy verification methods. Theory and algorithms*. Springer, 2014
- [23] G. I. Marchuk and V. V. Shaidurov. *Difference methods and their extrapolations*. Applications of Mathematics, New York: Springer, 1983.
- [24] I.V. Oseledets. *Approximation of $2^d \times 2^d$ matrices using tensor decomposition*. SIAM J. Matrix Anal. Appl., 31(4):2130-2145, 2010.
- [25] A. Ostrowski. Les estimations des erreurs a posteriori dans les procédés itératifs, C. R. Acad. Sci, Paris, Sér. AB 275 (1972), pp. A275A278.
- [26] S. Repin. *A posteriori error estimation for variational problems with uniformly convex functionals*, *Math. Comput.*, 69(2000), 230, 481–500.
- [27] S. Repin. *A Posteriori Estimates for Partial Differential Equations*. Walter de Gruyter, Berlin, 2008.

- [28] S. Repin, T. Samrowski, and S. Sauter. A posteriori error majorants of the modeling errors for elliptic homogenization problems. *C. R. Math. Acad. Sci. Paris* 351 (2013), no. 23-24, 877-882
- [29] S. Repin, T. Samrowski, and S. Sauter. Combined a posteriori modeling-discretization error estimate for elliptic problems with complicated interfaces. *ESAIM Math. Model. Numer. Anal.*, 46 (2012), no. 6, 1389-1405.
- [30] S. Repin, S. Sauter, A. Smolinski. A posteriori estimation of dimension reduction errors for elliptic problems in thin domains *SIAM J. Numer. Anal.*, 42(2004), No. 4, 1435-1451.
- [31] U. Schollwöck. *The density-matrix renormalization group in the age of matrix product states*, *Ann. Phys.* 326 (1) (2011) 96-192.
- [32] E. Zeidler. *Nonlinear functional analysis and its applications. I. Fixed-point theorems*, Springer-Verlag, New York, 1986.